Addressing Gender Bias in English to Spanish Translation with mBART

Ana María Zapata
UC Berkeley
anazapata@berkeley.edu

Abstract

In this paper we explore the evaluation and mitigation of gender bias in English to Spanish translation in mBART. Our approach leverages experiments developed by Saunders & Byrne, 2020 where they propose that a small dataset for data augmentation techniques can improve gender accuracy in translation tasks in transformers. Furthermore, we combine this dataset with a Google Wikipedia biographies dataset to not only improve the gender accuracy of the model but is fluency. Lastly, We try to analyze these translation models in the general translation metrics previously describe, while also understanding the limitations in the Spanish language.

1. Introduction

Detecting and raising awareness of the presence of gender bias in machine translation, mainly when we deal with gendered languages, is an important problem as learned models exhibit evidence of instantiated social biases and often replicate harmful associations in their training data (Caliskan et al., 2017; Chang et al, 2019, Quin et al 2022). For example, "gender bias manifests itself in training data which features more examples of men than of women. Tools trained on such data will then exhibit or amplify the biases (Zhao et al., 2017) and their harmful stereotypes" (Saunders & Byrne, 2020, 7724). Identifying and mitigating gender bias is an emerging and critical field for ensuring inclusion and accurate translation.

As previous research has shown, the evaluation and reduction of gender bias is a constant challenge in machine translation. In 2019 Stanovsky et. al presented the first challenge set for the analysis of gender bias in machine translation. Here, they establish that translations tend to be better for sentences involving the male gender, and for sentences containing stereotypical gender roles. (Saunders et al., 2020). For example, mentions of a nurse and a mechanic are translated into their stereotypical gender form. Meaning, the nurse would be translated into a female nurse in Spanish ("enfermera" and not "enfermero") even though is morphologically being identified as a male nurse. And the mechanic would be translated into a male mechanic even though is morphologically identified as a female mechanic ("mecánico" and not "mecánica").

To meet this challenge this project proposes first to evaluate an "out-of-the-box" mBART model pre-train to generate translations from English to Spanish. We selected mBART due to its intended translation and multilingual task. It is a “sequence-to-sequence demolishing auto encoder pre-trained on large-scale monolingual corpora in many languages using the BART objective”. Furthermore, we performed a series of data augmentation tasks following the experiment proposed by Saunders et al 2020 and we propose an extension of this experiment with the inclusion of a dataset that aims to create more context for the model.

Most previous works that follow winoMT only follow their metrics. But there is substantial variance in performance in the Spanish language in terms of the accuracy of nouns if we take into consideration ambiguous and epicene nouns. Therefore, in our experiment, we perform replications of the accuracy metrics in WinoMT, revise and update the occupation dictionary in Spanish taking into account all of the occupations contained in WinoMT[1] and analyze the translation with human evaluations to take into consideration epicene nouns.

2. Background

Several attempts have been developed to the evaluation and mitigation of gender bias in translation from English to gendered languages (Spanish, French, German, and several more). Escudé Font & Costa-jussà, 2019, analyzes the integration of two GloVe embeddings techniques in a transformer translation architecture. Vanmassenhove et al. (2018) tagged sentences with information on the speaker's gender, which affects the grammatical gender of words in some languages improving the translation quality of sentences spoken by women. Saunders et al. (2020) also used data augmentation techniques to mitigate gender bias in a transformer model. They improved the overall accuracy of the model with a small augmentation dataset. For languages with grammatical gender, data-centric approaches may be more difficult to apply, as gender may be represented not only in pronouns or occupations but also through the inflection of nouns, adjectives, and other parts of speech (Zmigrod et al., 2019).

3. Methods

Following the WinoMT challenge and evaluation, the problem we explore in this project is the error in the translation alignment of occupations in translations from English to Spanish. In the following example, we can see what Stanovsky et al. described as “a stereotypical assignment of gender roles changed the meaning of the translated sentence” by changing the occupation gender from female in English to male in Spanish.

reference	The developer argued with the designer because she did not like the design.
Incorrect Spanish translation	El desarrollador argumentó con el diseñador porque no le gustaba el diseño.
Correct Spanish translation	La desarrolladora argumentó con el diseñador porque no le gustaba el diseño.

[1] The original dictionary can be found here. And the revised version in the project github repository.

Figure 1. An example of gender bias is where we see how our baseline translation model interprets the phrase incorrectly. In the English sentence the “developer” is linked with the pronoun “she” identifying as a female developer. Nevertheless, the translation model assigns the developer as a male developer in “El desarrollador” and not the female form “La desarrolladora”.

Our intuition for mitigating gender bias follows the one stated by Saunders et al. (2020) where we believe that, taking into consideration the adaptive nature of a transformer model, we might see improvement in gender bias with the finetuning on a small dataset instead of a large one.

Baseline Model

We first consider the performance of a baseline mBART model fine-tuned to translate from English to Spanish. Specifically, we chose mBART for its primary task in translation and because it follows the “sequence-to-sequence denoising auto-encoder pre-trained on large-scale monolingual corpora in many languages using the BART objective”. Given this large-scale pre-training mBART has shown strengths in being fine-tuned to new target languages and custom and augmentation datasets since it presents powerful generalization and adaptation abilities to languages and domains that are not in the pre-trained corpora (Zihan et al., 2021, 2707).

Augmentation Experiments

We conducted experiments on English to-Spanish translation following the hypothesis developed by Saunders et al 2020. Here they constructed a “tiny, trivial set of gender-balanced English sentences”. In total, we have 388 Handcrafted occupation sentences in a gender-balance set. According to Saunders, the sentences follow the template:

The [OCCUPATION] finished his/her work.

As we previously stated one of the questions we would like to explore is “How much data do we really need to mitigate gender bias in a transformer model?”. Thus we would like to explore Saunders's hypothesis that the absence of gender bias can be treated as a small domain for the purposes of a model adaptation. In this case, we would like to explore how well a finetuned mBART model adapts to gender bias when is fine-tuned to a small domain dataset than attempting to fine-tune to a large dataset. Our second experiment takes Saunders's second dataset which is a set where the sentences that contain professions included in WinoMT are removed. (Saunders & Byrne, 2020, 7727) In total, we have again 388 sentences with balanced adjective-based sentences:

The tall [man/woman] finished [his/her] work.

For the exploration of the trade-off of gender accuracy and fluency and quality of the translations, our final experiment leverages the Translated Wikipedia Biographies dataset. The dataset is initially constructed to analyze common gender errors in machine translation. But since it offers instances with long text translations of 8 to 15 connected sentences referring to that central subject, we hypothesize that It could offer offers a layer of context when we combined them with Saunders's small and direct phrases[2]. Thus improving the quality of the translation as a whole.

Marie Curie was born in Warsaw. **She** was the first person to receive two Nobel Prizes in different specialties.	Marie Curie nació en Varsovia. Fue la primera persona en recibir dos premios Nobel en distintas especialidades.

Figure 2. Example of a phrase found in the Google dataset. Here we see, in comparison to Saunders's dataset, longer sentences that not only include a person and a noun but more context to the phrase and to the translation model.

All models are finetuned on a series of Google Colaboratory notebooks following the base parameters given by HuggingFace for the mBART translation tasks. The experiments use the "facebook/mbart-large-50-many-to-many-mmt" Model and its "MBart50Tokenizer". The source lang is set up for "es-XX" and the target language for "es-XX". The Weight decay (L2 regularization) coefficient is set at 0.01 to prevent overfitting during training. During fine-tuning training is continued without setting up a learning rate and with only one epoch. This allows fast training and a translation time in WinoMT of less than forty minutes.

Wino-MT challenge set and Evaluation Metrics

WinoMT presents a fairly balanced dataset that allows us to measure the model in terms of gender accuracy and how it performs when we have stereotypical occupations. For example, occupations like doctor, mechanic, or director are usually related to their male form, and nurse and teacher to their female form.

WinoMT	
Male	1826
Female	1822
Neutral	240
Total	3888

Figure 3. Details of WinoMT dataset.

Following the metrics stipulated in the WinoMT evaluation, where they evaluate first the grammatical gender of the primary entity from each translation hypothesis with automatic word alignment and morphological analysis, we evaluated our experiments with the following metrics:

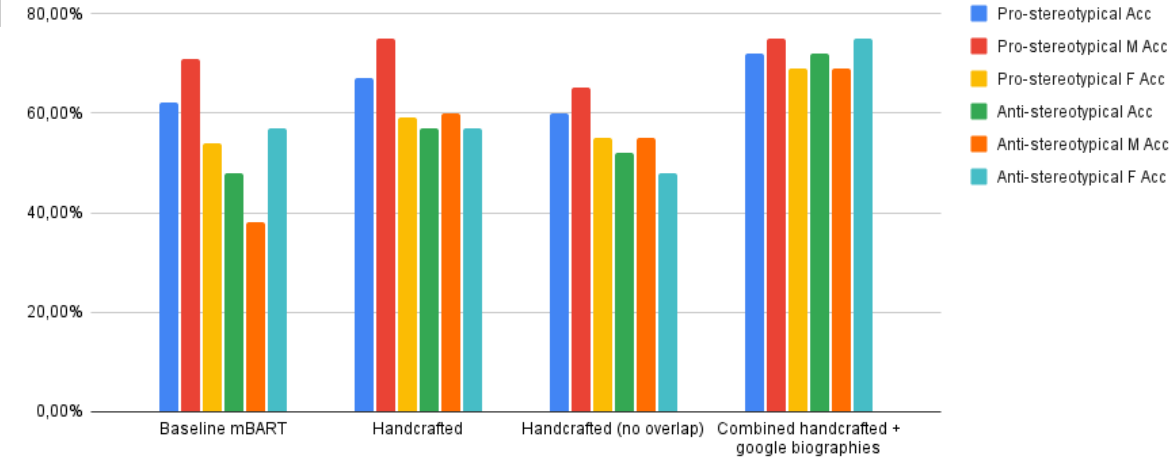
1. **Acc- Accuracy**: percentage of phrases with the correctly gendered primary entity. For this project, and following WinoMT, we took the phrases where the target entity is in index 1 and evaluate whether it contained the correct Spanish noun for female and male.
2. **G Acc- Gender Accuracy**: percentage of phrases with the correctly gendered occupation.

[2] More information related to the Translated Wikipedia Biographies by Google can be found here. As mentioned in the text, we understand that this dataset is intended as an evaluation of gender bias of translation models, but we would like to explore how well the mBART model fine-tuned in this dataset performs.

	Acc	G Acc	M Acc	F Acc
Baseline mBART	0.65	0.58	0.66	0.48
Handcrafted	0.81	0.66	0.70	0.60
Handcrafted (no overlap)	0.80	0.58	0.62	0.53
Combined handcrafted + google biographies	0.78	0.63	0.68	0.56

(up) Table 1. Accuracy of our experiments taking into consideration the metrics previously described. We evaluate here the complete WinoMT corpus and provide the evaluation for their anti and pro-test set in the following figure.

(down) Accuracy results for the pro and anti-stereotypical datasets tested in the four models.



3. **M Acc- Male Gender Accuracy**: percentage of phrases with the correctly gendered occupation.
4. **F Acc- Female Gender Accuracy**: percentage of phrases with the correctly gendered occupation.

We wish to reduce gender bias without reducing translation performances. Since WinoMT lacks references translation to perform metrics such as BLEU and BLEURT, we randomly selected a small set of phrases from each experiment to be evaluated by a human reviewer fluent in Spanish and English.

4. Results

Table 1 shows the results for our baseline model and our three selected experiments. In general, we notice that the four accuracy measures tend to increase with our data augmentations indicating that the transformer can identify a better gender accuracy once it has seen phrases that disrupt occupational stereotypes and include more inclusive approximations of occupations. For example. The inclusion o phrases where we see both a male and a female surgeon. This also proves that mBART is able to perform better when it is fine-tuned to a small domain dataset. Our best-performing model is our first experiment (Accuracy: 81%, Gender Accuracy: 66%, Male Accuracy: 70%, and Female Accuracy of 60%), the one where we only used the 388 phrase sets. Of these phrases, a total of 272 were the training pairs, 58 validation pairs, and 58 test pairs.

Even though this model is the overall best performer in terms of gender accuracy a closer look at the translations demonstrates that in some cases it's performing poorly in translation accuracy and fluency.

The model finetuned on the combination of Saunders's domain adaptation dataset (a total of 776 phrases) and the Google biographies (1472) came in second (Accuracy: 78%, Gender Accuracy: 63%, Male Accuracy: 68%, and Female Accuracy of 56%). These results not only keep confirming that the transformer adapts well to a small dataset of domain adaptations, but as we will later explore this model presents the best fluency and has the best translation accuracy in the human evaluation. The Handcrafted (no overlap) model came in third (Accuracy: 80%, Gender Accuracy: 58%, Male Accuracy: 62%, and Female Accuracy of 53%). These are excellent results that demonstrate mBART's ability to adapt to not only the dataset that has very direct occupation phrases but its ability to adapt to a small dataset that contains balanced adjective-based sentences. The results for accuracy of around 80% are at least 10% better than the Spanish metrics in the Saunders paper. The pro and anti-stereotypical results show that overall, the experiments perform better on both datasets, where we see a better performance in our combined model. This demonstrates that this model is the overall best performer (Gender Accuracy: 73%, Male Accuracy: 75%, and Female Accuracy of 69% in the pro stereotypical dataset and Gender Accuracy: 72%, Male Accuracy: 69%, and Female Accuracy of 75% in the anti stereotypical dataset. Showing an improvement of at least 10% in terms of our baseline model). We could attribute it to the inclusion of the dataset that shows more context to our model.

Human Evaluation

We also estimated the accuracy in gender and the accuracy in translation fluency by randomly sampling a small set of phrases for each experiment. We found that the combination of handcrafted and the Google biographies perform better in terms of fluency. In the following example, we can see that the first model correctly identifies the female gender for the first person and occupation "la jefa" but makes mistakes in adding a "ra" to the word "casa" and in the word for "mesada" it incorrectly describes a "menada". In the second phrase (our second model), we see that it incorrectly translates the gender of the first pronoun and the gendered version of the word for chief since it includes the male version. Finally, the last sentences demonstrate the overall better fluency in translation for our last model. Where we have a correct form for the first pronoun and occupation and better fluency in the phrase:

La jefa de la casara le dio una menada porque estaba satisfecha.

El jefe de familia dio una tipificación a la casara porque estaba satisfecha.

La directora le dio una tipificación a la alguacil porque estaba satisfecha.

After this evaluation, we can see that the augmentation dataset that contains sentences with enriched context contributed significantly in terms of translation fluency, without harming the increase of gender accuracy.

Upon looking at some of the incorrect predictions, the handcrafted models produce false translations for some occupations. For example, in some cases instead of the correct translation for the female form of lawyer, we found "abogadoa" instead of "abogada". This might be related to the fact that the model is only adding a letter "a" to some of the occupations. Upon looking at the results for the last model we see that his problem is less recurrent and that the contextual augmentation gives the model a better understanding of the words in Spanish. Conversely, the models don't produce Spanish translations for words that are difficult to translate into Spanish. For example, the word "mover" is just translated into "mover" since the term lacks a precise translation into Spanish and is often referred to as "persona de la mudanza", a person that moves things.

Finally, is important to note that the metrics developed here might lack an evaluation performance in epicene and ambiguous nouns, meaning nouns that in Spanish have the same form but can be used with masculine and feminine determiners. For example, the word for student (estudiante) is the same in the female and male versions in the Spanish translation, and it's differentiated in some cases by its pronouns or by the context of the phrase. ("la estudiante", "una estudiante" for the female version "el estudiante", "un estudiante" for the male version.

Conclusion

This project reaffirmed some of the findings of the work presented by Saunder et al. (2020) Transformers models are able to adapt to a small gender occupation domain dataset to perform better on the accuracy of gender in the WinoMT challenge. The project also explored how the model responds to the inclusion of a dataset that contains phrases with more context and that for better performance, not only in gender accuracy but in translation fluency this might be the better approach. Our project also demonstrated that we don't need large datasets to mitigate bias in transformers pre-train in translation tasks such as mBART. In future works, we would like to further explore gender bias mitigation techniques that are able to understand some of the specificity of the Spanish language in terms of gender accuracy, for example, epicene and ambiguous nouns in multilingual transformer models.

References

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing Gender Bias in Neural Machine Translation with Word Embeddings Techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy. Association for Computational Linguistics.

Oskar Van Der Wal, Jaap Jumelet, Katrin Schulz, and Willem Zuidema. 2022. The Birth of Bias: A case study on the evolution of gender bias in an English language model. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 75–75, Seattle, Washington. Association for Computational Linguistics.

Jun Wang, Benjamin Rubinstein, and Trevor Cohn. 2022. Measuring and Mitigating Name Biases in Neural Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2576–2590, Dublin, Ireland. Association for Computational Linguistics.

Fleisig, E., & Fellbaum, C. (2022). Mitigating Gender Bias in Machine Translation through Adversarial Learning. *arXiv preprint arXiv:2203.10675*.

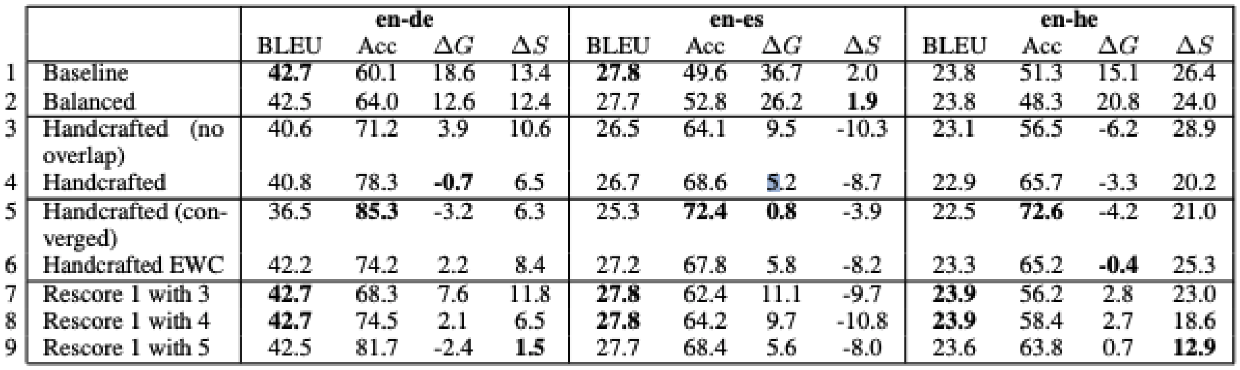
Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating Gender Bias in Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Danielle Saunders and Bill Byrne. 2020. Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

Zihan Liu, Genta Indra Winata, and Pascale Fung. 2021. Continual Mixed-Language Pre-Training for Extremely Low-Resource Neural Machine Translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2706–2718, Online. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

Appendix



		en-de				en-es				en-he			
		BLEU	Acc	ΔG	ΔS	BLEU	Acc	ΔG	ΔS	BLEU	Acc	ΔG	ΔS
1	Baseline	**42.7**	60.1	18.6	13.4	**27.8**	49.6	36.7	2.0	23.8	51.3	15.1	26.4
2	Balanced	42.5	64.0	12.6	12.4	27.7	52.8	26.2	**1.9**	23.8	48.3	20.8	24.0
3	Handcrafted (no overlap)	40.6	71.2	3.9	10.6	26.5	64.1	9.5	-10.3	23.1	56.5	-6.2	28.9
4	Handcrafted	40.8	78.3	**-0.7**	6.5	26.7	68.6	5.2	-8.7	22.9	65.7	-3.3	20.2
5	Handcrafted (converged)	36.5	**85.3**	-3.2	6.3	25.3	**72.4**	**0.8**	-3.9	22.5	**72.6**	-4.2	21.0
6	Handcrafted EWC	42.2	74.2	2.2	8.4	27.2	67.8	5.8	-8.2	23.3	65.2	**-0.4**	25.3
7	Rescore 1 with 3	**42.7**	68.3	7.6	11.8	**27.8**	62.4	11.1	-9.7	**23.9**	56.2	2.8	23.0
8	Rescore 1 with 4	**42.7**	74.5	2.1	6.5	**27.8**	64.2	9.7	-10.8	**23.9**	58.4	2.7	18.6
9	Rescore 1 with 5	42.5	81.7	-2.4	**1.5**	27.7	68.4	5.6	-8.0	23.6	63.8	0.7	**12.9**

General test set BLEU and WinoMT scores after fine-tuning on the handcrafted profession set, scores are quoted directly from (Saunders & Byrne, 2020, 7731)

	pro occupations			anti occupations		
	Gender accuracy	Male accuracy	Female accuracy	Gender accuracy	Male accuracy	Female accuracy
Baseline mBART	0.62	0.71	0.54	0.48	0.38	0.57
Handcrafted	0.67	0.75	0.59	0.57	0.60	0.57
Handcrafted (no overlap)	0.60	0.65	0.55	0.52	0.55	0.48
Combined handcrafted + google biographies	0.72	0.75	0.69	0.72	0.69	0.75

Accuracy results for the pro and anti-stereotypical datasets tested in the four models.